Educational Sciences: Theory & Practice • 14(2) • 973-976

www.edam.com.tr/estp
DOI: 10.12738/estp.2014.3.2107

Validity Evidence in Scale Development: The Application of Cross Validation and Classification-Sequencing Validation

Tülin ACAR[a]

Abstract

In literature, it has been observed that many enhanced criteria are limited by factor analysis techniques. Besides examinations of statistical structure and/or psychological structure, such validity studies as cross validation and classification-sequencing studies should be performed frequently. The purpose of this study is to examine cross validation and sequencing-classification validation at the same time with regard to two sub-samplings from an attitude scale concerning paranormal belief developed to guide researchers in interpreting its results. When the literature regarding scale development is taken into account, most of the scales have been developed in accordance with exploratory factor analysis. Even if the factor loads, model data conformity index, and the internal consistency reliability coefficients of the measuring devices are proper, the validity of the analysis should be examined through different methods. Parameter values which test as appropriate in the examined method may be found to be inappropriate or have different clues from other analyses. Therefore, the researcher's scale development should follow validity evidences through different methods.

Key Words

Classification and Sequencing Validation, Cross Validation, Double Consistency Index.

The validity of measuring devices used in education is one of the most important topics of the measuring device development process. Validity concept is a criterion for the fact that it serves as a measuring device (Croker & Algina, 2008; Downing & Haladyna, 2006; Kane, 2006). In other words, identifying the degree of an expected structure and of an observed structure is the structural validity of a test (Baykal, 1994). Thus, the validity of a measurement is directly proportionate to the purpose being measured by the device. Therefore, validity is not a concept to be considered independent of purpose and therefore a set of evidences should be collected.

Validity approach according to the purpose of measurement is generally discussed in 3 groups: content, criteria and structural validity (Brualdi, 1999; Erkuş, 2003; Hopkins, 1998). Content validity is related to the fact that the items to be tested represent the structure to be measured. In criterion supported validity, the relationship between points from one test and points from another test are taken as criteria to be examined. Structural validity is the degree to which significant organizational or psychological structures are represented.

The validity of measuring devices, test items, and accordingly the measurements used in education is one of the basic problems with the impartiality of

a **Tülin ACAR, Ph.D.**, is an Educational Measurement and Evaluation specialist. Research interests include hierarchical linear models, differential item functioning, psychometric properties of tests, educational statistics, and multivariate statistical analysis. *Correspondence:* Parantez Education, Research Publisher, Selanik Street No: 46/4 Kızılay-Çankaya, Ankara, Turkey. Email: totbicer@gmail.com

measurement areas. As is known, one of the primary purposes of measurement applications in education is to obtain information about individuals or test items. Therefore, flawless measurement devices/ results are required. The validity of a measurement devices' results should be high. However, one of the factors which affect validity negatively is a "biased" item. The fact that a test includes biased items will undoubtedly destroy an evaluations' credibility and limit its ability to be carried out in accordance with the results of the test. The impartiality of items is detected through a set of psychometric procedures in accordance with the test theory (Camilli & Shepard, 1994; Holland & Wainer, 1993; Millsap & Everson, 1993; Raju & Ellis, 2002; Zumbo, 1999).

Stuck (1995), in his study, proposed that especially both measurement mistakes and biased items are among the factors which destroy a structure's validity. Validity problem is a degree of sufficiency, therefore he proposed feasibility validity instead of construct feasibility.

According to Messick (1995), in educational and psychological measurements, six distinguishable features were emphasized for validity: content, substance, structure, ability to generalize, externalization and consequence validity. All these features have been evaluated as evidence for collecting information to validate a study.

In order to identify the "construct" validity of a measurement device, factor analysis is applied for a validity study (Croncbach & Meehl, 1955). As is known, grouping dependent on the correlation of the points observed is carried out. This grouping is related to the items within the factor analysis measuring device. Thus, structure(s) in which related items gravitate to measuring may come into being. However, factor analysis is discussed as "exploratory factor analysis" and "confirmatory factor analysis" in itself (Pohlmann, 2004; Stapleton, 1997)

Groupings dependent on the correlation concerning the scoring of items are classified as "exploratory factor analysis." Therefore, the constructs to be put forth together with "exploratory factor analysis" is also called "statistical constructs" in some sources (Knight, 2000; Pohlmann, 2004; Stapleton, 1997). In confirmatory factor analysis, item-construct relations based on theory are tested instead of the scores of the items. Thus, in confirmatory factor analysis, the construct to be approached is also called a "psychological construct" (Knight, 2000; Pohlmann, 2004).

Guilford, who termed construct validity, factorial validity or validity concepts for the first time 60 years ago stated that the answer to the question: "Does a test measure a desired expected construct?" is a type of validity problem and this validity problem can be solved through the factor analysis method (Stapleton, 1997). Today, however, concepts such as validity proofs, correlation between measurements, internal consistency, reliability coefficient, validity distinction, cross validation, classification validity, and sequencing validity are examined.

Purpose and Importance of Research

In literature, it has been observed that many enhanced criteria are limited to factor analysis techniques. Besides the examination of statistical and/or psychological structures, validity studies such as cross validation and classification-sequencing studies should be frequently included.

The purpose of this study is to examine cross validation and sequencing-classification validity at the same time with regard to two sub-samplings from an attitude scale concerning paranormal belief which was developed in order to guide researchers in interpreting the results.

Method

Study Group

The revelation of attitudes and beliefs occurs between the ages of 18-21 (Hökelekli, 1998, p. 280). The study group of this research consists of 947 people above the age of 18.

Data Collection Tool

For a data collection tool, the final norm tool which was developed in order to measure attitudes concerning paranormal beliefs has been used. This tool consists of 23 items. The final form has been scaled according to the quinary Likert type. In the development process of the scale, a pool attitude scale consisting of 70 items was prepared according to the development principles (see Tezbaşaran, 1997) and this was then applied to a group of 100 people. As a result of this application the difference between the sub-group and superior group averages and the significance of these differences was determined by the t test. For construct validity, exploratory factor analysis used the Cronbach alpha coefficient, and the total material correlations for the internal consistency of materials were

evaluated. The final form has been constructed. In the second phase, the final form was applied to 947 people and the attitude features of people's paranormal beliefs were subjected to exploratory factor analysis. The representation levels of latent variable were also evaluated. The psychometric features and results concerning structure have been included in the researcher's study which is called "Efforts on Measuring Attitudes Regarding Paranormal Beliefs."

Data Analysis Techniques

Cross validation is the investigation of the fact that the competency of a model in two or more random samples taken from the same population is invariable. Thus, in this study, multiple group features of the LISREL package program have been used to evaluate whether a measured psychological structure has cross validation.

The non-existence hypothesis for cross validation states that the measurement model parameters (factor loads, factor variances, factor covariances and measuring error variances) between two samples need to be identical (invariable).

Set instructions for the second sample indicates that factor variance, co-variance and measuring error variances are different between two samples.

For classification and sequencing validity, the Double Consistency Index from Erkuş (2003) was used for calculations. Development of validity using this method is as follows: the test materials are separated into two sides as single and double. In the two sides, the total points for each individual sample are contained. These score totals are arranged in order from highest to lowest value. After that, the match between sub-groups and superior groups in both sides is evaluated to be 27%. In the event that the test carries out a consistent classification (in other words, it distinguishes consistently), use of the double consistency index depends on the fact that individuals classified in the superior group from the first half of the test stay in that group for the second half; and that individuals classified in the sub-group from the first half of the test stay in the sub-group for the second half. In both halves of the test, through frequency differences in superior and sub groups (27%), an index increase in value of 0.00 and 1.00 was achieved. When an index value draws close to 0.00, it states inconsistent classification and when it draws close to 1.00 it states consistent classification.

Results

Findings Concerning Cross Validation

The Alpha internal consistency coefficient for the 23-item scale sampled from 947 people was calculated at 0.824. Split-half consistency (consistency between the forms consisting of the first twelve and the last eleven items) has been calculated at 0.803 and the Gutlam split-half consistency coefficient has been calculated at 0.656 and the correlation between the forms consisting of single and even numbered items has been calculated at 0.656. It may be said that the responses to the scale item show consistency and determination.

The Cronbach Alpha internal consistency coefficient of the first sample separated at random was calculated at .817. The Cronbach Alpha consistency coefficient of the second sample was calculated at .830. Thus, the data from both samples were found to have similar internal consistency coefficients.

In order to evaluate the cross validity of two separate samples' measuring model, the chi-square difference test was used.

The Chi square difference test measures the difference between conformity of the chi square tests for the measuring models only under the non-existence and alternative hypotheses. The degree of freedom is the difference between the measuring model's degree of freedom only under the nonexistence and alternative hypotheses.

Significance levels of 0.299 and 0.499 were calculated respectively for the chi square difference test and the measuring model parameters (factor loads, factor variances, factor co-variance and measuring error variances). This shows that the levels are invariable. In other words, cross validity of the measuring model for scale item is supported in both samples.

The resemblance rate and chi square statistical equation for the first sample is $X^2(506)=1807.26$, $p<0.01$, where the root mean square error approach(RMSEA) = 0.091. The resemblance rate and chi square statistical equation for the second sample is $X^2(483)=1781.21$, $p<0.01$, where the root mean square error approach (RMSEA) = 0.093. In both samples, the standardized root mean square residual(S-RMR) = 0.07; the comparative fit index (CFI)= 0.88, the goodness of fit index (GFI)= 0.82, the normed fit index (NFI)= 0.84, and the relative fit index (RFI)= 0.84.

It can be stated that, as a result of confirmatory factor analysis, the single-factor structure of the scale provides acceptable and valid results.

Findings regarding Classification and Sequencing Validity

After the scale with 23 items is divided into two halves, the total score for individuals concerning the scale items is obtained. Individuals are listed according to their score totals for both halves. A group rate of 27% was chosen by beginning from the highest point listed in descending order. This first group is called the superior group. Then, proceeding down the list, individuals are formed into sub-groups consisting of odd numbers. Superior groups are formed consisting of even numbers and the individuals are placed into sub-groups and superior groups. In subsequent proceedings, the points of individuals are no longer taken into account. According to the double consistency calculation formula with regard to the rate of 27% the number of individuals in the sub-groups and superior groups is 256. The number of people taking place in both odd and even numbered forms in the sub-groups is 72. The number of people taking place in both odd and even numbered forms in the superior group is 160. According to the calculation formula, the obtained frequencies are calculated at 0.45. It can be stated that according to the index varying between 0.00 and 1.00, the sequencing-classification validity of 0.45 can be considered middle level.

Discussion

When the literature regarding scale development is taken into account, most of the scales have been developed in accordance with exploratory factor analysis. Moreover, many scales have been used only once, for the development of purpose. To summarize, what is left turns to scale rubbish.

Scale developing is a process, and in this process it is required that items are regulated again, that calculated factorial statistics are renewed, and that different samples are tested. In the scale development process, the generally examined structure (an implicit feature) is finalized in the article study. Undoubtedly, these kinds of studies should be discussed in more than one article.

Even if the factor loads, the model data conformity index, and the internal consistency reliability coefficients of measuring devices are proper, validity analysis should be examined through different methods. Parameter values detected appropriate via the examined method may be found to be inappropriate or to have different clues. Therefore, a researcher who is developing a scale should follow validity evidences through many different methods.

References/Kaynakça

Baykal, A. (1994). Davranışların ölçülmesinde yapısal geçerlilik göstergesi. *Türk Psikoloji Dergisi*, *33*, 45-50.

Brualdi, A. (1999). *Traditional and modern concepts of validity*. Retrieved from http://eric.ed.gov/PDFS/ED435714.pdf

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Erkuş, A. (2003). *Psikometri üzerine yazılar.* Ankara: Türk Psikologlar Derneği Yayınları.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston: Allyn and Bacon.

Hökelekli, H. (1998). *Din psikolojisi*. Ankara: Türkiye Diyanet Vakfı Yayınları.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

Knight, J. L. (2000, November). *Toward reflective judgment in exploratory factor analysis decisions: Determining the extraction method and number of factors to retain*. Paper presented at the Annual Meeting of the Mid-South Educational research Associations, Bowling Green, KY. (ERIC Document No. ED 449224)

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Millsap, R. E., & Everson, H. T. (1993). Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.

Pohlmann, J. T. (2004). Use and interpretation of faktor analysis in the journal of educational research: 1992-2002. *ProQuest Psychology Journals, 98*(1), 14-22.

Raju, N. S., & Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 156-188). San Francisco, CA: Jossey-Bass.

Stapleton, C. D. (1997). *Basic concepts and procedures of confirmatory factor analysis.* Educational Research Association, Reports-Evaluative (142), Speeches / Meeting Papers (150).

Stuck, I. (1995, April). *Heresies of the new unified notion of test validity*. Paper presented at the Meeting of the National Council on Measurement in Education, San Francisco, CA.

Tezbaşaran, A. A. (1997). *Likert tipi ölçek geliştirme kılavuzu*. Ankara: Türk Psikologlar Derneği.

Zumbo, B. D. (1999). *A handbook on the theory and methods of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for Binary and Likert-Type (Ordinal) item scores.* Retrieved from http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf